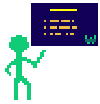
**Création et analyse des profils phylogénétiques construits à partir du génome de 8 souches de la bactérie *Escherichia coli***

Pour rappel :  un profil phylogénétique décrit la présence ou l'absence d'un gène (d'une protéine) dans un ensemble de génomes comparé à un génome de référence. Ces gènes (ou protéines) doivent présenter une relation d'orthologie avec le gène (la protéine) du génome de référence.   
  
**Comment pratiquement et rapidement peut-on identifier les gènes orthologues entre deux espèces ?**La technique la plus couramment utilisée consiste à analyser les résultats d'un programme appelé Blast qui permet de comparer rapidement une séquence donnée (sequencerequête ou query) à un ensemble de séquences constituant une banque. Le résultat du Blast est un alignement local indiquant les régions présentant des similarités entre les deux séquences. Cette similarité est quantifiée par le pourcentage d'identité trouvé entre les deux séquences alignées, le score de l'alignement (qui a été normalisé et est appelé bit score) et la e-value qui correspond au nombre attendu d’alignement qui par chance aurait un score >= au score obtenu avec nos deux "vraies" séquences. Ce nombre  est évalué en regardant les scores des alignements générés par comparaison de séquences aléatoires ayant même longueur et même composition que la séquence requête. Plus la e-value est faible plus l'alignement obtenu est significatif.  
  
Pour rechercher les gènes orthologues, la comparaison se fait souvent au niveau de leurs produits protéiques. Nous utiliserons donc ici, par abus de langage, les termes d'orthologues et de paralogues pour des séquences protéiques, ceci pour simplifier l'écriture.  
  
Dans les fichiers joints vous trouverez donc le résultat de la comparaison de chaque protéine d'un génome (chaque séquence sera à son tour séquence requête) avec l'ensemble des protéines d'un autre génome constituant donc la banque de séquences. Le programme utilisé pour la comparaison est BlastP (version du blast pour comparer une séquence protéique à une banque protéique).  
  
Nomenclature utilisée et exemple   
Code à 5 lettres se référant au nom de l'espèce suivi d'une cinquième lettre assignée à la souche : EcolA:  Ecol pour *Escherichia coli* et la dernière lettre se référant à la souche (A pour K12 MG1655, B pour O157:H7, D pour CFT073, L pour K12 DH10B,  M pour SECEC SMS-3-5, R pour IAII, U pour UMN026 et X pour BL21)  
EcolB\_EcolD.blt : chaque séquence du génome de la souche EcolB01 sera une séquence query qui sera comparée à l'ensemble des séquences protéiques de la souche EcolD01. Comme résultats on ne gardera que le premier hit (donc la séquence de EcolD01 qui possède la plus petite e-value). On parlera alors de **Best Hit.**Si on prend l'exemple de la séquence EcolB01.ECS002son best hit dans EcolD01 est la séquence EcolD01.THRA. On notera BH(EcolB01.ECS002, EcolD01.THRA) la e-value est de 0 et le score de 1590.  
EcolD\_EcolB.blt : chaque séquence du génome de la souche EcolD01 sera une séquence query qui sera comparée à l'ensemble des séquences protéiques de la souche EcolB0101. Comme résultats on ne gardera que le premier hit.  
Si on prend l'exemple de la séquence EcolD01.THRA, son best hit est la séquence EcolB01.ECS002. On a donc BH(EcolD01.THRA,EcolB01.ECS002)  
  
Dans cet exemple on a donc BH(EcolB01.ECS002, EcolD01.THRA) = BH(EcolD01.THRA,EcolB01.ECS002), les deux séquences sont réciproquement best hit l'une de l'autre. **Dans ce cas on parle de Best Bidirectional Hits appelé dans la littérature BBH.   
  
La première condition pour que deux protéines soient orthologues est qu'elles soient BBH. Ensuite, pour que cette relation soit 1:1, il ne faut pas qu'une des deux séquences possède une séquence paralogue dont le score d'alignement serait supérieur à celui du BBH.**Nous devons donc vérifier qu'il n'y a pas eu de duplication du gène dans un des deux génomes après leur divergence (qui serait donc plus récente que la divergence) et qui résulterait dans la présence d'une séquence paralogue (dans le même génome) ayant un score de d'alignement de Blast supérieur à celui obtenu entre les séquences BBH de nos deux génomes. Pour cela, pour chaque génome, chacune de ces séquences protéiques a été comparée à l'ensemble des autres séquences protéiques et le premier et donc meilleur hit obtenu est appelé Best Paralog, noté BP.  
Les résultats de ces comparaisons sont trouvés dans les fichiers portant deux fois le nom de la souche.  
Si nous reprenons l'exemple précédant :  
EcolB\_EcolB.blt : chaque séquence protéique de la souche EcolB01 est comparée à toutes les autres protéines de cette souche.  
Le best hit de la séquence EcolB01.ECS002 est EcolB01.ECS4869 après le premier hit trivial qui est celui de la séquence requête avec elle-même.  Ces deux séquences sont donc best paralog et le score de l'alignement obtenu est 341.  
EcolD\_EcolD.blt : chaque séquence protéique de la souche EcolD01 est comparée à toutes les autres protéines de cette souche. Le best hit de la séquence EcolD01.THRA est EcolD01.METL. Ces deux séquences sont donc best paralog et le score de l'alignement obtenu est 344.  
  
  
Les deux scores BP(EcolB01.ECS002,EcolB01.ECS4869) = 341 et BP(EcolD01.THRA,EcolD01.METL) = 344 < BBH((EcolB01.ECS002, EcolD01.THRA) = 1590  
  
**Il n'existe donc pas de séquences paralogues dans aucun des deux génomes possédant un score d'alignement supérieur au BBH. Donc nous pouvons conclure que EcolB01.ECS002 et EcolD01.THRA sont deux séquences orthologues 1:1.**  
 **Les données fournies :  ([archive ci-jointe](http://silico.biotoul.fr/enseignement/m1-mabs/BGPG/projet/genomes_coli.tar.gz))**

**8 souches ont été analysées : ecolA, EcolB, EcolD, EcolL,EcolM, EcolR, EcolU et EcolX.**

Pour chacune d'entre elle il y a 8 fichiers : Comparaison de chaque protéine de la souche avec les autres protéines de la même souche (fichier ou le nom de la souche apparaît deux fois), comparaison de chaque protéine de la souche avec l'ensemble des protéines des 7 autres souches (dans le nom du fichier, le nom de la première souche est celle dont les séquences sont utilisées comme query et le deuxième est le nom de la souche dont les séquences ont servi à constituer la banque).  
Dans chaque cas seul le best hit est reporté.  
Le fichier contient le résultat du BlastP sous forme de tableau avec les colonnes correspondant aux données suivantes:  
  
nom séquence query , nom séquence banque, % d'identité, longueur de l'alignement, nombre de substitutions, nombre de gaps (indels), position début alignement séquence query, position fin alignement séquence query, position début alignement séquence banque, position fin alignement séquence banque, e-value, score normalisé (bit score).  
  
**Le projet demandé**Permettre à l'utilisateur de choisir un génome de référence parmi les 8 fournis. Par comparaison de chacun des sept génomes restant avec le génome de référence, identifier les protéines orthologues1:1. A partir de ces résultats construire les profils phylogénétiques en codant 1 la présence de la protéine orthologue et 0 son absence. A partir de ce tableau, construire une matrice de pseudo-distances entre les profils phylogénétiques en utilisant l'indice de Jaccard. On aura pris soin avant de construire la matrice de pseudo-distances d'éliminer les résultats triviaux (protéines orthologues présentes dans tous les génomes). Appliquer ensuite une méthode de classification automatique pour trouver les clusters de profils similaires. Analyser deux clusters pris au choix au niveau fonctionnel (les protéines appartiennent-elles a une même processus biologique?).  
  
Pour cela, il faudra programmer l'identification des protéines orthologues, la construction des profils phylogénétiques. Vous utiliserez ensuite R pour calculer la matrice de pseudo-distances et réaliser la classification automatique. Pour l'analyse fonctionnelle des clusters sélectionnés vous pourrez vous aider de la recherche d'annotations sur-représentées (programme sook de mobyle disponible à l'adresse <http://silico.biotoul.fr/cgi-bin/mobyle/portal.py>). Pour le trouver aller dans annotation puis enrichment. Les annotations Gene Ontology ainsi que les mots-clés SwissProt sont disponibles pour la souche EcolA (K-12/MG1655). Pour les exploiter, il faudra, bien évidemment, que les gènes du cluster à analyser aient des orthologues chez cette souche.  
Rendre un rapport décrivant votre démarche et intégrant l'algorithme du(des) programme(s) ainsi que la description de vos résultats. Le code du  (des) programme(s) sera envoyé séparément.